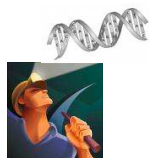
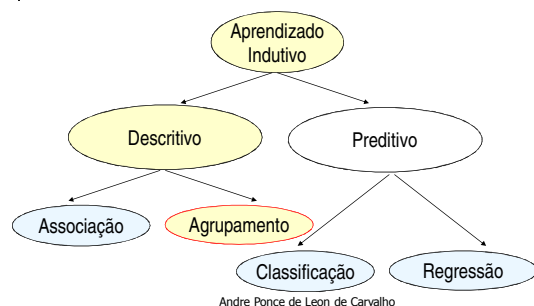
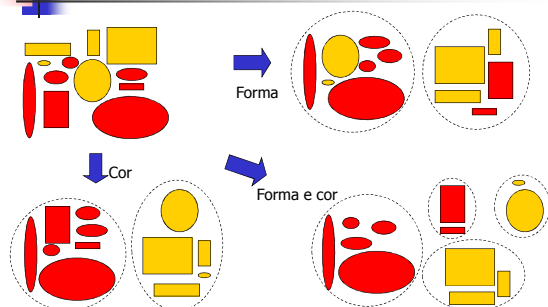
Mineração de Dados em Biologia Molecular
Agrupamento de Dados
André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Tópicos
Agrupamento de dados
Análise de cluster
Dificuldades em agrupamento
Algoritmos de agrupamento
Validação
Aplicações
Andre Ponce de Leon de Carvalho
2

Agrupamento
Aprendizado Indutivo
Descritivo
Preditivo
Associação
Agrupamento
Classificação
Regressão
Andre Ponce de Leon de Carvalho
3

Agrupamento
Organização de um conjunto de objetos em grupos (clusters)
De acordo com alguma forma de similaridade ou relação entre eles
Como organizar?
Andre Ponce de Leon de Carvalho
4

Agrupamento
Forma
Cor
Forma e cor
Andre Ponce de Leon de Carvalho
5

Objetivo
Encontrar conjunto de clusters que maximizam SM ou minimizam DM
SM: medida de similaridade
DM: medida de dissimilaridade
Quanto maior a homogeneidade dentro dos grupos e a diferença entre os grupos, melhor
Alternativas
Busca exaustiva
Técnicas mais sofisticadas
Andre Ponce de Leon de Carvalho
6

Agrupamento de dados
x1
x2
Algoritmo de agrupamento
x1
x2
Andre Ponce de Leon de Carvalho
7

Quantos clusters?
Dados originais
6 clusters
2 clusters
4 clusters
Andre Ponce de Leon de Carvalho
8

Diferentes alternativas
Possíveis formatos
Compacto
Alongado
Elipsoidal
Espiral
Andre Ponce de Leon de Carvalho
9

Agrupamento de dados
Definição do que é um agrupamento
Imprecisa
Depende de:
Natureza dos dados
Resultados desejados
Existem várias
Andre Ponce de Leon de Carvalho
10

Tipos de agrupamento
Seja $X = \{x_1, x_2, \ldots, x_n\}$ o conjunto de todos os dados
Tarefa: colocar cada X_i em um dos m clusters $C_1, C_2, \ldots, C_m$
Agrupamentos podem ser de dois tipos:
Tipo 1: duro (crisp)
Tipo 2: fuzzy
Andre Ponce de Leon de Carvalho
11

Tipos de agrupamento
Agrupamento crisp
Cada objeto X_i pertence ou não a cada cluster C_j
$C_i \neq \emptyset, i = 1, \ldots, m$ $\bigcup_{i=1}^{m} C_i = X$
$C_i \cap C_j = \emptyset, i \neq j, i, j \in \{1,2, \ldots, m\}$
Objeto em C_i mais semelhante a outros objetos em C_i do que àqueles em C_j, $i \neq j$
Andre Ponce de Leon de Carvalho
12

Tipos de agrupamento

- Agrupamento fuzzy
  - Usa uma função de pertinência para definir o quanto um elemento pertence a um grupo

$\mu_j : X \rightarrow [0, 1]$

$$\sum_{j=1}^{m} \mu_j(x_i) = 1, i \in \{1, ..., n\}$$

m = número de grupos
n = número de objetos

$$0 < \sum_{i=1}^{n} \mu_j(x_i) < n, \; j \in \{1, ..., m\}$$

Algoritmos de agrupamento

- Busca exaustiva
  - Tentar todos os possíveis agrupamentos de tamanho *m* (para vários valores de *m)*
  - Números de Stirling do segundo tipo
    - Número de formas de particionar *n* dados em *m* subconjuntos não vazios

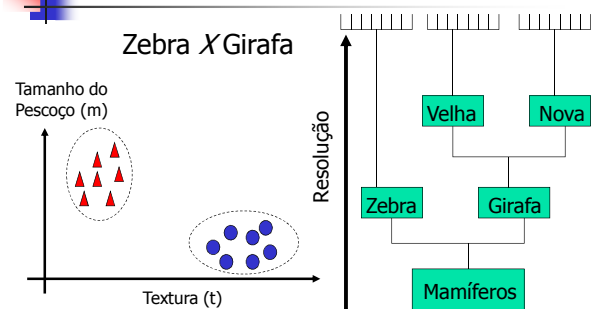
$$>> \binom{n}{m} \geq \left(\frac{n}{m}\right)^m$$

m = número de grupos
n = número de objetos

  - Impraticável

Algoritmos de agrupamento

- Particionais
  - Protótipos (erro quadrático médio)
  - Densidade
- Hierárquicos
- Baseados em grids
- Baseados em grafos
- Outros algoritmos
  - Ex.: Redes neurais SOM

Particional X Hierárquico

Algoritmos particionais

- Principais características
  - Produzem um único agrupamento
  - A maioria utiliza abordagem "gulosa" (*greedy*)
    - Sempre procura escolher melhor alternativa atual, sem considerar conseqüências futuras
    - Uma vez tomada uma decisão, ela não é mais alterada
    - Geralmente resultado depende da ordem de apresentação dos exemplos

Algoritmos particionais

- K-médias
- K-médias ótimo
- K-médias seqüencial
- SOM
- FCM
- DENCLUE
- CLICK
- CAST
- SNN

Algoritmo k-médias

- Supor n objetos $x_1, x_2, ..., x_n$ a serem agrupados em k clusters, k < n
  - Seja μ_i a média dos objetos do cluster C_i
  - Medida de distância pode ser utilizada para definir a que cluster um objeto pertence
    - $\mathbf{x_p} \in$ cluster C_i se $d(\mathbf{x_p}, \mu_i)$ é menor que todas as k–1 distâncias entre $\mathbf{x_p}$ e μ_j, j = 1, 2, ..., k e $i \neq j$

Algoritmo k-médias

1 Sugerir médias $\mu_1, \mu_2, ..., \mu_k$ iniciais
2 Repetir
Usar as médias sugeridas para agrupar os objetos em K clusters
Para i variando de 1 a K
Substituir μ_i pela média de todos os objetos do cluster C_i
Até nenhuma das médias mudar

Algoritmo k-médias

- Médias iniciais
  - Exemplos (vetores) aleatórios
  - Elementos aleatoriamente escolhidos do conjunto de treinamento

Limitações do k-médias

- Escolha do valor de K
- K-médias tem problemas quando os grupos têm:
  - Diferentes densidades
  - Formatos não hiper-esféricos
- Tem problemas também quando os dados contêm *outliers*

Exercício

- Agrupar, utilizando k-médias, os dados abaixo em 2 grupos:
  - X_1 = 1, 0, 1, 1
  - X_2 = 1, 0, 0, 0
  - X_3 = 0, 1, 1, 0
  - X_4 = 1, 1, 1, 1
  - X_5 = 0, 0, 0 ,1

Validação de agrupamentos

- Existem várias medidas para avaliar qualidade de classificadores
  - Acurácia, precisão, revocação, F1
- Como avaliar os clusters gerados por um algoritmo de agrupamento?

Medidas de validação

- Existem várias medidas de validação
  - Julgam aspectos diferentes
- Podem ser divididas em três grupos
  - Índices ou critérios externos
    - Medem o quanto os rótulos dos grupos casam com a classe verdadeira
  - Índices ou critérios internos
    - Medem a qualidade da partição obtida sem considerar informações externas
  - Índices ou critérios relativos
    - Usados para comparar duas partições ou grupos

Medidas internas

- Coesão de clusters
  - Mede o quão relacionados estão os objetos dentro de um cluster
- Separação de clusters
  - Mede o quão distinto ou separado cada cluster é dos demais clusters
- Silhueta

Silhueta

- Combina coesão com separação
- Calculada para cada objeto que faz parte de um agrupamento
  - Baseada na proximidade entre os objetos de um cluster e na distância dos objetos de um cluster ao cluster mais próximo

Silhueta

- Para cada objeto x_i
  - $a(x_i)$: distância média de x_i aos outros objetos de seu cluster
  - $b(x_i)$: min (distância média de x_i a todos os objetos de cada outro cluster)

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & \text{se } a(x_i) < b(x_i) \\ 0, & \text{se } a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1, & \text{se } a(x_i) > b(x_i) \end{cases}$$

- Largura média da silhueta
  - Média sobre todos os objetos do conjunto de dados
  - Valor entre -1 e 1 (quanto mais próximo de 1, melhor)

Algoritmos hierárquicos

- Utilizam diagrama de árvore (dendograma)
  - Produzem uma sequência (hierarquia) de agrupamentos
- Historicamente utilizados em áreas que utilizam estrutura hierárquica de dados
  - Ex.: Biologia e arqueologia

Algoritmos hierárquicos

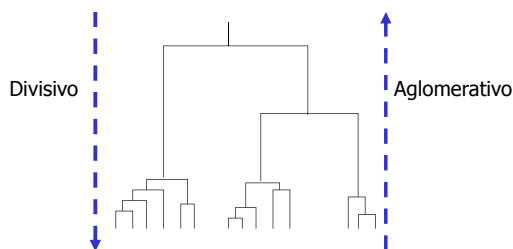
- Conceito de representação hierárquica de dados originou-se na Biologia
  - Algoritmos de agrupamento hierárquicos ≅ estrutura hierárquica da taxonomia de Linnaean
  - Biólogos geralmente preferem agrupamentos hierárquicos

Algoritmos hierárquicos

- Aplicações na biologia geralmente não se preocupam com o número ótimo de clusters
  - Biólogos geralmente estão interessados na hierarquia completa

Algoritmos hierárquicos

- Podem ser de dois tipos:
  - Aglomerativos: combinam, repetidamente, dois clusters em um
    - A cada passo, combina os dois clusters mais próximos
  - Divisivos: Dividem, repetidamente, um cluster em dois
    - A cada passo, divide o cluster menos homogêneo em dois novos clusters

Exemplo

Algoritmos aglomerativos

- Começam com $C_0 = \{\{x_1\}, ..., \{x_n\}\}$
- A cada passo t, combinam dois clusters em um, produzindo:
  - $|C_{t+1}| = |C_t| - 1$ e $C_t \subset C_{t+1}$
- No passo final (passo n-1) tem-se a hierarquia:
  - $C_0 = \{\{x_1\}, ..., \{x_n\}\} \subset C_1 ... \subset C_{n-1} = \{x_1, ..., x_n\}$

Algoritmos divisivos

- Começam com $C_0 = \{x_1, ..., x_n\}$
- A cada passo t, dividem um cluster em dois, produzindo:
  - $|C_{t+1}| = |C_t| + 1$ e $C_{t+1} \subset C_t$
- No passo final (passo n-1) tem-se a hierarquia:
  - $C_{n-1} = \{\{x_1\}, ..., \{x_n\}\} \subset ... \subset C_0 = \{x_1, ..., x_n\}$

Algoritmo aglomerativo

1 Incializar $C_0 = \{\{x_1\}, ... ,\{x_n\}\}$
2 Para t = 1 até n – 1 faça
Encontrar o par de clusters mais próximos (C_i, C_j)
$C_t = (C_{t-1} - \{C_i, C_j\}) \cup \{\{C_i \cup C_j\}\}$ / atualizar centros*

Algoritmos hierárquicos

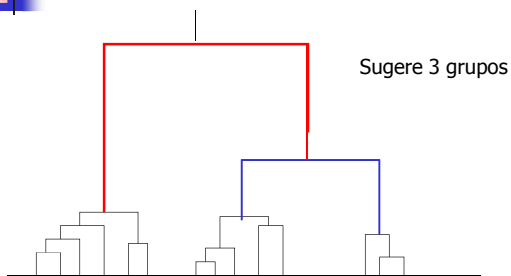
- Existe uma grande variedade de algoritmos hierárquicos
  - Geralmente diferem na forma de calcular distância inter-clusters

$$d_{AB} = \min_{\substack{i \in A \\ j \in B}} (d_{ij})$$ Por ligação simples (single-link)

$$d_{AB} = \max_{\substack{i \in A \\ j \in B}} (d_{ij})$$ Por ligação completa (complete-link)

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$ Pela média do grupo (average-link)

Algoritmos hierárquicos

- Como escolher uma partição?
  - Partição com n clusters
    - Selecionando partição com *n* clusters na seqüência de agrupamentos da hierarquia
  - Partição que melhor se encaixa nos dados
    - Procurar no dendograma grandes mudanças em níveis adjacentes
      - Nesse caso, uma mudança de *j* para *j-1* grupos pode indicar que *j* é o melhor número de grupos
      - Existem outros procedimentos, alguns mais objetivos

Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício

- Agrupar os dados em dois grupos usando o algoritmo K-médias e medida de silhueta
  - Usar k = 2 e k = 3
  - Informação sobre a classe não deve ser usada
- Em que grupos seriam colocados os novos casos?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Considerações finais

- Abordagens tradicionais de agrupamento são muito utilizadas em AM
  - Várias definições de agrupamento
  - Diversos algoritmos
- Dificuldade de validar agrupamentos encontrados
- Semi-supervisionado

Perguntas
Andre Ponce de Leon de Carvalho
43